HCUP Nationwide Inpatient Sample

Calculating Nationwide Inpatient Sample Variances, 2001

May 30, 2003

TABLE OF CONTENTS

INDEX OF TABLES

EXECUTIVE SUMMARY

The Healthcare Cost and Utilization Project (HCUP) is a Federal-State-Industry partnership to build a standardized, multi-State health data system. The 2001 NIS is a stratified sample of hospitals drawn from the subset of hospitals in 33 states that make their data available to the HCUP project and that can be matched to the AHA survey data. Hospitals are stratified by region, location/teaching status, bed size category, and ownership. The NIS includes all discharges from the sampled hospitals.

This document describes how to calculate simple statistics, including variances, from the 2001 Nationwide Inpatient Sample (NIS) taking into account the sampling design and sample discharge weights. It contains the program code required to calculate sample totals, means, rates, and their variances with three commonly used statistical programming languages that run on personal computers: SAS, SUDAAN and Stata. This report also provides results of example calculations from all three statistical packages using the 2001 NIS. These analyses provide a baseline against which users can compare their own estimates to ensure their programming accuracy. They also demonstrate that the results are virtually the same for all three statistical packages. Finally, we discuss alternative concepts of variance and other methods that could be applied to calculate variances.

INTRODUCTION

The Healthcare Cost and Utilization Project (HCUP) is a Federal-State-Industry partnership to build a standardized, multi-State health data system. In September 2000, the Agency for Healthcare Research and Quality (AHRQ) provided funding for The MEDSTAT Group, Inc. (MEDSTAT) to continue existing development efforts and to expand this health data system through data year 2003. The major goals of this expansion are increasing the number of states contributing inpatient data, expanding the ambulatory surgery and emergency department databases, and possibly adding an ambulatory care database. One objective, already achieved, was a redesign of the Nationwide Inpatient Sample (NIS) sampling and weighting strategy.

This document describes how to calculate statistics, including variances, from the 2001 Nationwide Inpatient Sample (NIS) taking into account the sampling design and sample discharge weights. It gives the program code required to calculate sample totals, means, rates, and their variances with three commonly used statistical programming languages that run on personal computers:

1) SAS Version 8.02,
2) SUDAAN Release 8.02 (SAS-callable standalone version), and
3) Stata SE Version 8.0.

All three languages have procedures for calculating sample statistics and appropriate variances based on data from complex sampling designs. This is important, because unweighted statistics and analyses that fail to account for the NIS sample design could yield biased estimates. Although this report does not cover multivariate statistical procedures like regression analysis, some concepts introduced in this report carry over to those areas of analysis, as well.

Several statistical programming packages allow weighted analyses. If the user prefers to use a statistical package other than these three, it is likely that the options and statements for that package will be similar to those for one of the three packages covered by this report. For an excellent review of such programs, visit the following web site: http://www.fas.harvard.edu/~stats/survey-soft/survey-soft.html. Appendix A contains a summary of survey analysis capabilities for SAS, Stata, and SUDAAN copied from this web site.

This report also gives the results of example calculations from all three statistical packages. Therefore, the user can run the program code in this document and check the results obtained against the results reported here.

This introduction continues below with a brief overview of the NIS sample design and the discharge weights that accompany the NIS database. The user desiring a more comprehensive account should refer to the final report on the NIS sampling and weighting strategies[1]. This introduction ends with a brief discussion on the treatment of missing values in the data.

NIS Sample Design

The final sample design is as follows. The hospital universe is defined by all hospitals that were open during any part of the calendar year and were designated as community hospitals in the

[1] *Changes in the NIS Sampling and Weighting Strategy for 1998, Final Report*. January 18, 2002. Agency for Healthcare Research and Quality, Rockville, MD.

2001 American Hospital Association (AHA) Annual Survey of Hospitals, excluding rehabilitation hospitals. For purposes of the NIS, the definition of a community hospital is that used by the AHA: "all nonfederal short-term general and other specialty hospitals, excluding hospital units of institutions." Consequently, Veterans Hospitals and other federal hospitals are also excluded. The NIS is a stratified sample of hospitals drawn from the subset of hospitals in states that make their data available to the HCUP project and *that can be matched to the AHA survey data.* There are 60 strata. Hospitals are stratified by region, location/teaching status (within region), bed size category (within region and location/teaching status), and ownership (within region, location/teaching, and bed size categories). The regions are defined by the four census regions (NE, NC, S, and W). Location is defined by AHA's designation of urban or rural. Teaching hospitals are those with membership in the Council of Teaching Hospitals (COTH), or with an AMA-approved residency program, or with an intern-to-bed ratio of 25 percent or higher. Bed size categories are small, medium, and large, with separate size cut points defined for each combination of hospital region, teaching status, and urban/rural designation. Ownership breakdowns are based on the degree of observed ownership variation within each region across bed size categories. Within each stratum, we draw a systematic random sample of hospitals equal in size to 20 percent of the universe for that stratum. The hospitals were sorted by the first three digits of their zip code for the systematic sample. The NIS includes all discharges from the sampled hospitals. For more details, see the report, *Design of the HCUP Nationwide Inpatient Sample, 2001.* This report is available on the 2001 NIS Documentation CD-ROM and on the HCUP User Support Website at www.hcup-us.ahrq.gov.

NIS Sample Weights

The discharge sample weights are calculated within each sampling stratum as the ratio of discharges in the universe to discharges in the sample. Consequently, the discharge sample weights are constant for all discharges within each stratum with the exception of adjustments for hospitals with missing quarters of data. The number of discharges in the universe is calculated from the total number of discharges reported in the 2001 AHA hospital survey data for non-rehabilitation community hospitals. Therefore, the sum of the sample weights in each stratum represents the total number of discharges reported in the AHA survey.

In the 2001 NIS files, the discharge weight data element is named DISCWT. To produce national estimates we use DISCWT to weight sampled discharges in the NIS to the discharges from all non-rehabilitation community hospitals located in the U.S.[2]

[2] Note: For the 2000 NIS, DISCWT should be used to create national estimates for all analyses except those that involve total charges. The data element DISCWTCHARGE should be used to create national estimates of total charges. Texas discharges were not included in the calculation of DISCWTCHARGE because total charges were not available for the first half of 2000 from that state. Consequently, in the 2000 NIS DISCWTCHARGE differs from DISCWT for NIS hospitals in the South region.

Missing Values

The procedures presented in this report omit cases with missing values from all calculations. Missing values for any reason can compromise the quality of estimates. If the outcome for discharges with missing values is different from the outcome for discharges with valid values, then sample estimates for that outcome will be biased and will not accurately represent the discharge population. There are several techniques available to help overcome this bias. One strategy is to use imputation to replace missing values with acceptable values. Another strategy is to use sample weight adjustments to compensate for missing values[3]. This data preparation and adjustment is outside the scope of this report. However, if necessary, it should be done before analyzing data with the statistical procedures presented here.

On the other hand, if the cases with and without missing values are assumed to be similar with respect to their outcomes, then no adjustment may be necessary for estimates of means and rates because the means and rates based on nonmissing cases would be representative of the means and rates of missing cases. However, some adjustment may still be necessary for the estimates of totals. Totals (of non-negative variables) would tend to be underestimated in the presence of missing values of the variable for which the total is estimated because the cases with missing values would be omitted from the calculations.

The next section establishes some sampling concepts in a short discussion of a formula that could be used to calculate the variance of a total from the NIS sample. The following sections contain the program code required to estimate some sample statistics and their variances using each of the three statistical packages. We demonstrate that the results are identical or very similar for all of the programs. Finally, we discuss the finite population correction, alternative concepts of variance, and other methods that could be applied to calculate variances.

RATIONALE AND FORMULAS FOR NIS VARIANCE CALCULATIONS

For a simple random sample of discharges, the usual variance calculations are appropriate. For example, the unbiased estimate for the variance of hospital length of stay (LOS) based on a sample of n discharges would be calculated as:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

where x_i is the LOS for discharge i, and $\bar{x}$ is the mean LOS over the sample of n discharges. Consequently, the estimated standard error of the mean would be calculated as: $\hat{\sigma}/\sqrt{n}$.

However, the sample of NIS discharges is not a simple random sample. First, we selected a stratified sample of hospitals. The hospital sampling rate within each stratum was approximately 20 percent of the hospital universe. We then selected all hospital discharges from each of the sampled hospitals. Consequently, the NIS sample resembles a stratified single-stage cluster sample with hospitals as the clusters sampled in the first stage at the rate of about 20 percent and with discharges as the elementary units sampled at a rate of 100 percent.

A complication, which we ignore in calculating variances below, is that the hospital sampling frame did not contain the entire universe of U.S. hospitals. The frame contained only hospitals

[3] See, for example, Foreman, E.K., *Survey Sampling Principles*, Dekker, New York, 1991, Chapter 10.

in the 33 states for which all-payer discharge data were made available to the HCUP project. To the extent that states in the frame differ from other states on outcomes within each stratum, this could lead to biased estimates. Consequently, users should compare estimates from the NIS to other benchmarks whenever they are available. For the 2000 NIS, we compared a broad range of estimates from the NIS to estimates from the National Hospital Discharge Survey (NHDS) and the Medicare MedPAR file[4]. Most estimates were consistent among the three data sources. The updated report for the current NIS will be available on the HCUP User Support Website (www.hcup-us.ahrq.gov) later in the year of data release (e.g., the 2001 *NIS Comparison Report* will be available in fall of 2003).

We will consider the NIS sample as a two-stage cluster sample, even though the sampling rate was 100 percent at the second stage, because the analyst may wish to consider the sampling rate at less than 100 percent when considering how to handle missing values or when using one of the two 10-percent subsamples. The variance formula for a stratified two-stage cluster sample employs weights and components for the two stages of sampling. This is necessary to account for the possibility that sample discharges within hospitals may be more homogeneous in their outcomes than sample discharges between hospitals. If the analyst wants finite population estimates, then factors are also needed to correct for the proportion of the universe included in the sample at each level (finite population correction factors).

The following example is meant to illustrate the "behind the scenes" calculations that statistical programs make for variances based on sample designs. The reader may safely move on to the next section of this report without understanding the technical details of this example.

For this example, consider the estimated total of a variable Y, calculated as the weighted sum:

$$T = \sum_s \sum_h \sum_d w_{shd}\, y_{shd}$$

where:

y_{shd} = the observed value of variable Y for sample discharge *d* within sample hospital *h* within stratum *s*.

w_{shd} = a set of discharge weights or any other constants over the set of sample discharges, hospitals, and strata. The NIS sample weights are constant for discharges within each hospital. However, we retain the subscript *d* to account for the possibility that an analyst would want to adjust the weights to account for missing values in a way that creates unequal weights across patients within a hospital. For example, the weights might be made to vary according to some patient-level characteristic, such as the patient's DRG, if the rate of missing values varies by that characteristic. The existing weights within strata sometimes differ by hospital to account for underreporting of discharges by some NIS hospitals.

In any case, an estimate of the variance of T from the sample is:

$$\hat{\sigma}_T^2 = \sum_s (1 - f_s) n_s V_s + \sum_s f_s \sum_h (1 - f_{sh}) n_{sh} V_{sh} \qquad (1)$$

where:

[4] 2000 NIS Comparison Report.

f_s = the proportion of the universe hospitals sampled in stratum *s*. In the NIS, this is usually close to 20 percent, but it varies because the number of universe hospitals is usually not an exact multiple of five within the strata. If we wish to generalize results to a broader set of hospitals and patients outside the 2001 hospital population, then we would set $f_s = 0$. This might be desirable, for example, if the analyst wishes to draw inferences about a future year or wishes to use the results to set policy going forward.

n_s = the number of hospitals within stratum s.

f_{sh} = the proportion of the discharges in the sample from sample hospital *h* within stratum *s*. For the NIS, $f_{sh} = 1$. However, we show this term because an analyst may wish to consider this a sample from an infinite population (of possible patients), in which case $f_{sh} = 0$, rather than a finite population.

n_{sh} = the number of discharges in hospital h within stratum s.

V_s = the component of variance due to the first stage of sampling (variation among hospitals within stratum s):

$$V_s = \frac{\sum_h \left(\sum_d w_{shd} y_{shd} - \frac{\sum_h \sum_d w_{shd} y_{shd}}{n_s} \right)^2}{(n_s - 1)}$$

Notice that the numerator is the sum of squared deviations of the individual hospital totals from the mean hospital total, and the sum is over all hospitals in stratum s, similar to the familiar calculation for the variance of any sample statistic. Also notice in equation 1 that this term is multiplied by zero if $f_s = 1$. In that case, all hospitals within stratum s are sampled, and the estimated total for that stratum has no sampling error associated with it.

V_{sh} = the component of variance due to the second stage of sampling (variation among discharges within hospital h in stratum s):

$$V_{sh} = \frac{\sum_d \left(w_{shd} y_{shd} - \frac{\sum_d w_{shd} y_{shd}}{n_{sh}} \right)^2}{n_{sh} - 1}$$

Again, this calculation of a variance is familiar. The numerator is the sum of squared deviations of the individual weighted discharge totals from the mean weighted discharge total for each hospital h in stratum s. If the sampling rate $f_{sh} = 1$ for hospital h in stratum s, then this term is multiplied by zero in equation 1 because the hospital total is estimated without error

Many statistical packages use variance formulas similar to (1) to estimate variances for simple statistics such as means and totals.

It is important to recognize that these variance calculations assume that the analyst is interested in making inferences to the finite population of 2001 hospital discharges. As the sampling fraction *f* approaches 1, the sampling variance approaches zero. If the analyst is interested in making inferences to another population, not the specific discharges represented in the 2001 discharge population, then the sampling fraction *f* should be set to zero. Our examples will not use the finite population correction (fpc). However, we will indicate the effect of the fpc and how the fpc could be incorporated.

EXAMPLES OF NIS VARIANCE CALCULATIONS

The example dataset is a subset of the NIS created by selecting all records with a Clinical Classifications Software (CCS) diagnosis category code equal to 50: diabetes mellitus with complications. Clinical Classifications Software (CCS) is a tool for clustering patient diagnoses and procedures into a manageable number of clinically meaningful categories developed at the Agency for Healthcare Research and Quality (*Clinical Classifications Software*. Fact Sheet. Agency for Healthcare Research and Quality, Rockville, MD. http://www.ahrq.gov/data/hcup/ccsfact.htm).

To obtain estimates, we created an analysis file from the NIS using SAS to select the subset of discharges with complicated diabetes for the analysis (DXCSS1 = 50). For Stata we also generated a data file in ASCII format. We used the SAS-callable version of the SUDAAN software, so we were able to use the SAS file for analysis by SUDAAN. However, like Stata, the stand-alone version of SUDAAN would also require the ASCII format. The SAS, SUDAAN, and Stata program code for these steps are shown below, along with examples of the output produced by each program.

In these examples, the following conventions apply:

- Lowercase words denote NIS variable names.
- UPPERCASE WORDS denote keywords and options that are part of the programming language.
- *Italicized words* denote information to be supplied by the researcher.
- **Bold words** denote comments.

SAS Programming Statements

```
/* Create analysis file */

LIBNAME In "location of NIS file" ;

DATA Diabetes ;
  SET  In.nis_2001_core;
  IF dxccs1 = 50 ;
  dischgs = 1 ;

/* Obtain estimates:  The following SAS code produces estimates of
the sums, the means, and the standard errors for the number of
discharges, the length of stay, and the total hospital charges */

PROC SURVEYMEANS DATA=Diabetes SUM STD MEAN STDERR ;
  WEIGHT discwt ;
  CLASS died ;
  CLUSTER hospid ;
  STRATA NIS_stratum ;
  VAR dischgs los died totchg ;

/* Note: If finite population estimates of standard errors are    */
/* wanted, then the PROC SURVEYMEANS statement could include the */
/* option "RATE = .20", which gives the approximate hospital      */
/* sampling rate in each stratum.                                 */
```

- The PROC SURVEYMEANS statement invokes the SAS procedure.
- The DATA= option requests that the analysis be performed on the file specified. If this statement is omitted, SAS uses the most recently created dataset.
- The SUM option requests the sum of discharges. The variable DISCHGS is set to equal 1 for every record, so its sum estimates the total number of discharges.
- The STD option requests the standard error of the sum.
- The MEAN and STDERR options request that the mean and its standard error be printed. The default statistics are the mean, its standard error, and 95% confidence limits.
- The WEIGHT statement weights each record by the value of the variable DISCWT.
- The CLASS statement identifies DIED as a categorical variable for which a ratio analysis is performed (ratio of sum of DIED to sum of DISCWT).
- The STRATUM statement specifies NIS_STRATUM as the stratum identifier.
- The CLUSTER statement specifies HOSPID as the cluster identifier.

- The VAR statement requests the statistics for the variables DISCHGS, LOS, TOTCHG and DIED. If the VAR statement is omitted, statistics will be calculated for all of the variables in the dataset except for those listed in the WEIGHT, STRATUM or CLUSTER statement.

These commands produced the following output:

SAS Output

Data Summary	
Number of Strata	60
Number of Clusters	967
Number of Observations	92426
Sum of Weights	461161.024

Class Level Information			
Class Variable	**Label**	**Levels**	**Values**
DIED	Died during hospitalization	2	1 2

Statistics					
Variable	**Label**	**Mean**	**Std Error of Mean**	**Sum**	**Std Dev**
Dischgs	Discharges	1.0000	0	461161	8965.1604
LOS	Length of stay (cleaned)	5.5830	0.0593	2574645	60392
TOTCHG	Total Charges (cleaned)	15917	431.8004	7249884071	251139678
DIED=0	Died during hospitalization	0.9861	0.000436	454352	8835.6602
DIED=1		0.0139	0.000436	6396.5554	238.9113

SUDAAN Programming Statements

```
/* The following code produces the estimate and standard error for
total hospital discharges, mean length of stay, and mean total
charges using the SAS-callable version of SUDAAN */

/* Create analysis file */

LIBNAME In "location of NIS file" ;

DATA Diabetes ;
  SET  In.nis_2001_core;
  IF dxccs1 = 50 ;
  dischgs = 1 ;
  died = died + 1 ; /* It is necessary to recode 0,1 variables */
                    /* for use by SUDAAN because it considers  */
                    /* zeroes as missing values in categorical */
                    /* variables                               */

PROC DESCRIPT DATA=Diabetes FILETYPE=SAS DESIGN=WR ;
  WEIGHT discwt ;
  NEST nis_stratum hospid ;
  SETENV colwidth = 24 ;
  VAR los dischgs totchg ;
  PRINT TOTAL SETOTAL MEAN SEMEAN ;

/* SUDAAN does not allow continuous and categorical variables  */
/* to be analyzed in a single step. The following procedure    */
/* calculates statistics for the categorical variable "died".  */

PROC DESCRIPT DATA=Diabetes FILETYPE=SAS DESIGN=WR ;
  WEIGHT discwt ;
  NEST nis_stratum hospid ;
  VAR died ;
  CATLEVEL 2 ;     /* This specifies the number of categories for died */
```

- The PROC DESCRIPT statement invokes the procedure.
- The DATA= option specifies the dataset name.
- The FILETYPE option specifies that this is a SAS file.
- The DESIGN = options identifies this as a sample With Replacement (WR). NIS hospitals were sampled without replacement. However, this specification is appropriate when the hospital "population" is considered very large or conceptually infinite.
- The WEIGHT statement identifies "discwt" as the weight variable.
- The SETENV statement increases the column width to allow the printing of numbers larger than the default width.

- The NEST statement identifies the first variable listed (nis_stratum), as the stratum variable and the second variable (hospid) as the primary sampling unit.
- The VAR statement lists the variables to be included in the analysis.

These statements produced the following output:

SUDAAN Output

```
Variance Estimation Method: Taylor Series (WR)

-----------------------------------------------------------------
| Variable        |                  |                          |
-----------------------------------------------------------------
| Length of stay  | Total            |               2574644.58 |
| (cleaned)       | SE Total         |                 60391.66 |
|                 | Mean             |                     5.58 |
|                 | SE Mean          |                     0.06 |
-----------------------------------------------------------------
| DISCHGS         | Total            |                461161.02 |
|                 | SE Total         |                  8965.06 |
|                 | Mean             |                     1.00 |
|                 | SE Mean          |                     0.00 |
-----------------------------------------------------------------
| Total charges   | Total            |            7249884071.11 |
| (cleaned)       | SE Total         |             253421414.47 |
|                 | Mean             |                 15916.89 |
|                 | SE Mean          |                   432.30 |
-----------------------------------------------------------------
Variance Estimation Method: Taylor Series (WR)

SUDAAN RATIO ESTIMATES
Variance Estimation Method: Taylor Series (WR)

-----------------------------------------------------
| Variable        |                  |             |
-----------------------------------------------------
| Died during     | Sample Size      |        92344 |
| hospitalization | Weighted Size    |    460748.77 |
|                 | Total            |      6396.56 |
|                 | Percent          |         1.39 |
|                 | SE Percent       |         0.04 |
-----------------------------------------------------
```

Stata Programming Statements

```
/* Using SAS, create an ASCII file for use by STATA */

LIBNAME In "location of NIS file" ;

DATA _NULL_ ;
  SET  In.nis_2001_core ;
  IF dxccs1 = 50 ;
  FILE  fileref ;
  IF los < 0 THEN los = . ;
  IF died < 0 THEN died = . ;
  IF totchg < 0 THEN totchg = . ;
  PUT nis_stratum 1-4 hospid 6-10 died 12 los 14-17
      dischgs 19 totchg 21-27 +1 discwt ;

/* Obtain STATA estimates */

SET MEMORY 32000
INFILE nis_stratum hospid died los dischgs totchg
       discwt USING "dataset name""
SVYSET [PWEIGHT = discwt], STRATA (nis_stratum), PSU (hospid)
SYVTOTAL dischgs
SVYMEAN los totchg
SVYRATIO died dischgs
```

- The SET MEMORY command increases the memory allocated to Stata to 32 megabytes. The default memory allocation of 1 MB probably will be too small for most subsets of the NIS and will need to be changed prior to any analyses. (This command works only for the Windows and Unix versions of Stata. Users of other versions should see the manual specific to their operating system).
- The INFILE command lists the variables to read in from the dataset created for this analysis.
- The SVYSET command identifies the weight variable, the stratification variable, and the primary sampling unit.
- The SYVTOTAL command requests the estimate of the total and standard error for the variable listed.
- The SVYMEAN command requests the estimate of the mean and its standard error for the variables listed.
- The SVYRATIO command requests the ratio of the two variables listed, in this case, of those who died to total discharges.

These commands produce the following output:

Stata Output

```
Survey total estimation

Pweight:  discwt                        Number of obs    =     92426
Strata:   nis_stratum                   Number of strata =        60
PSU:      hospid                        Number of PSUs   =       967
                                        Population size  = 461161.02

   Total   Estimate    Std. Err.    [95% Conf. Interval]     Deff

Dischgs    461161      8965.057          443566.4           478755.7


Survey mean estimation

Pweight:  discwt                        Number of obs    =     92426
Strata:   nis_stratum                   Number of strata =        60
PSU:      hospid                        Number of PSUs   =       967
                                        Population size  = 454345.05

   Mean    Estimate    Std. Err.    [95% Conf. Interval]     Deff

   Los     5.58         .0593           5.47   5.70          6.96
   Totchg  15916.89    431.80       15069.44   16764.34      25.98


Survey ratio estimation

Pweight:  discwt                        Number of obs    =     92344
Strata:   nis_stratum                   Number of strata =        60
PSU:      hospid                        Number of PSUs   =       967
                                        Population size  = 460748.77

   Ratio       Estimate  Std. Err.     [95% Conf. Interval]      Deff

died/dischgs   .013883   .0004359       .013028   .014738        1.28
```

Comparison of Estimates

Table 1 displays the estimates from each of the three statistical programming packages using the program code described earlier. The estimates are virtually identical.

Table 1: Comparison of SAS, SUDAAN and Stata Results,

Complicated Diabetes
Nationwide Inpatient Sample (NIS), 2001

VARIABLE (Standard Error)	SAS	SUDAAN	Stata
Total Discharges	461,161 (8,965)	461,161 (8,965)	461,161 (8,965)
In-hospital Mortality %	1.39 (.044)	1.39 (.04)	1.39 (.044)
Length of Stay	5.58 (.059)	5.58 (.06)	5.58 (.059)
Total Charges	$15,917 (432)	$15,917 (432)	$15,917 (432)

Finite Population Corrections

The NIS sample contains all discharges from about 20 percent of all hospitals nationwide. Therefore, analysts may want to "correct" variance estimates to account for the fact that sampling error is attributable only to the remaining 80 percent of the hospital population. Hence, the finite population correction factor (fpc), the multiple of the infinite population variance, is equal to about 80 percent. This means that the standard errors reported in the above table would be multiplied by 89.4 percent (the square-root of 80 percent). While this decreases the estimated standard error by a little over 10 percent, the fpc should be applied only when inferences are being made to the specific population of patients actually hospitalized during 2001. Usually analysts prefer not to use the fpc because they are interested in the long-run results for hospitals. For example, interest centers on the true, long-run mortality rate for a hospital rather than on the mortality rate actually observed in 2001.

Using the NIS Subsamples

To obtain variance estimates using data from the NIS 10 percent subsample, only one modification to the programs is needed. The WEIGHT variable must be changed to use the DISCWT10 variable, which is simply 10 times the DISCWT.

The results from SAS, SUDAAN and Stata using one of the 10 percent subsamples are shown in Table 2 below.

Table 2. Comparison of SAS, SUDAAN, and Stata Results

Complicated Diabetes
Using a 10 percent subsample of the
Nationwide Inpatient Sample (NIS), 2001

VARIABLE (Standard Error)	SAS	SUDAAN	Stata
Total Discharges	454,419 (9,529)	454,419 (9,529)	454,419 (9,529)
In-hospital Mortality %	1.63 (.13)	1.63 (.13)	1.63 (.13)
Length of Stay	5.55 (.09)	5.55 (.09)	5.55 (.09)
Total Charges	$15,968 (556)	$15,968 (556)	$15,968 (556)

DISCUSSION

Alternative Concepts of Variance

Sometimes analysts require variance calculations based on finite-sample theory. According to finite-sample theory, the intent of the estimation process is to obtain estimates that are precise representations of the nationwide population of patients treated during a specific year. In the context of the NIS, any estimates that attempt to accurately describe characteristics (such as expenditure and utilization patterns or hospital market factors) and interrelationships among characteristics of hospitals and discharges during a specific year from 1988 to 2001 should be governed by finite-sample theory.

Alternatively, in the study of hypothetical population outcomes not limited to a specific point in time, analysts may be less interested in specific characteristics from the finite population (and time period) from which the *sample* was drawn, than they are in hypothetical characteristics of a conceptual "superpopulation" from which any particular finite *population* in a given year might have been drawn. According to this superpopulation model, the nationwide population in a given year is only a snapshot in time of the possible interrelationships among hospital, market, and discharge characteristics. In a given year, all possible interactions between such characteristics may not have been observed, but analysts may wish to predict or simulate interrelationships that may occur in the future.

Under the finite-population model, the variances of estimates approach zero as the sampling fraction approaches one, since the population is defined at that point in time, and because the estimate is for a characteristic as it existed at the time of sampling. This is in contrast to the superpopulation model, which adopts a stochastic viewpoint rather than a deterministic viewpoint. That is, the nationwide population in a particular year is viewed as a random sample of some underlying superpopulation over time.

Estimation Techniques

Different methods are used for calculating variances under the two sample theories. Under the superpopulation (stochastic) model, procedures have been developed to draw inferences using weights from complex samples[5]. In this context, the survey weights are not used to weight the sampled cases to the universe, because the universe is conceptually infinite in size. Instead, these weights are used to produce unbiased estimates of parameters that govern the superpopulation.

In addition to the methods shown in this report, standard errors can be estimated by validation and cross-validation techniques. Given that a very large number of observations will be available for most analyses, it may be feasible to set aside a part of the data for validation purposes. Standard errors and confidence intervals can then be calculated from the validation data.

If the analytical file is too small to set aside a large validation sample, cross-validation techniques may be used. For example, tenfold cross-validation would split the data into ten equal-sized subsets. The estimation would take place in ten iterations. At each iteration, the outcome of interest is predicted for one-tenth of the observations by an estimate based on a model fit to the other nine-tenths of the observations. Unbiased estimates of error variance are then obtained by comparing the actual values to the predicted values obtained in this manner.

Longitudinal analyses of hospital-level outcomes may be biased if they are based on any subset of NIS hospitals limited to continuous NIS membership. Hospitals that continue in the NIS for multiple consecutive years are a subset of the hospitals in the NIS for any one of those years. In particular, such subsets would tend to contain fewer hospitals that opened, closed, split, merged, or changed strata. Further, the sample weights were developed as annual, cross-sectional weights rather than longitudinal weights. Therefore, different weights might be required, depending on the statistical methods employed by the analyst.

One approach to consider in hospital-level longitudinal analyses is to use repeated-measure models that allow hospitals to have missing values for some years. However, the data are not actually missing for some hospitals, such as those that closed during the study period. In any case, the analyses may be more efficient (e.g., produce more precise estimates) if they account for the potential correlation between repeated measures on the same hospital over time, yet incorporate data from all hospitals in the sample during the study period.

The two nonoverlapping 10 percent subsamples of discharges may be used to validate models and obtain unbiased estimates of standard errors. That is, one subsample may be used to estimate statistical models, and the other subsample may be used to test the fit of those models on new data. This is an important analytical step, particularly in exploratory studies, where one runs the risk of fitting noise.

For example, it is well known that the percentage of variance explained by a regression, R^2, is generally overestimated by the data used to fit a model. The regression model could be estimated from the first subsample and then applied to the second subsample. The squared correlation between the actual and predicted value in the second subsample is an unbiased estimate of the model's true explanatory power when applied to new data.

[5] Potthoff, R.F., M.A. Woodbury, and K.G. Manton, *"Equivalent Sample Size" and "Equivalent Degrees of Freedom" Refinements for Inference Using Survey Weights Under Superpopulation Models. Journal of the American Statistical Association*, Vol. 87, (1992), pp. 383-396.

CONCLUSIONS

We found that the three statistical packages produced identical or very similar values for weighted sample statistics, including sample variances. The new SAS procedures for the calculation of sample statistics from complex surveys should be a great convenience for SAS users. In prior versions of SAS, such calculations required either user-written programs or the data had to be imported and processed by other statistical packages.

APPENDIX A:
SUMMARY OF SURVEY ANALYSIS CAPABILITIES FOR SAS, STATA, AND SUDAAN[6]

Summary of survey software: SAS/STAT

Vendor

SAS Institute Inc.

Types of Designs That Can Be Accommodated

For the sample selection procedure, the sample design can be a complex multistage sample design that includes stratification, clustering, replication, and unequal probabilities of selection.

For survey data analysis procedures, the sample design can be a complex survey sample design with stratification, clustering, unequal weighting, and with or without replacement.

Types of Estimands and Statistical Analyses That Can Be Accommodated

SAS/STAT Software now provides the SURVEYSELECT, SURVEYMEANS, and SURVEYREG procedures. These procedures were made available as experimental procedures in Version 7 of the SAS System, and were released as production procedures in Version 8. (Future releases of SAS are intended to handle analyses of frequency data (scheduled for Release 9) and logistic regression (Release 9.1). The release of SAS that you have can make a big difference to the facilities in this area.)

- The **SURVEYSELECT procedure** provides a variety of methods for selecting probability-based random samples. The procedure can select a simple random sample, or samples with design features such as stratification, clustering or multistage sampling, or unequal probabilities of selection. It can accomodate very large sampling frames. It can draw a replicated sampling, i.e. a sample composed of a set of replicates, each selected in the same way.

  PROC SURVEYSELECT accepts the sampling frame as a SAS data set. Control language specifies the selection methods, the desired sample size or sampling rate, and other parameters. The output data set contains the selected units, with selection probabilities and sampling weights.

- The **SURVEYMEANS procedure** estimates population totals, means, and ratios (SAS 8.2 and later), with estimates of their variances, confidence limits, and other descriptive statistics, under sample designs that may include stratification, clustering, and unequal weighting.

- The **SURVEYREG procedure** estimates regression coefficients by generalized least squares, using elementwise regression, assuming that the regression coefficients are the same across strata and PSUs.

Restrictions on Number of Variables or Observations.

None

Primary Methods Used for Variance Estimation.

Taylor expansion.

[6] This information was copied from the following website http://www.fas.harvard.edu/~stats/survey-soft/survey-soft.html maintained by Harvard University and the Survey Research Methods Section of the American Statistical Association.

General Description of the "Feel" of the Software.

The interface is similar to other SAS procedures. Programs may be entered from command files or through a windowing system. The Explorer window is used to view and manage SAS files. The Program Editor is used to enter, edit, and submit SAS programs, and messages appear in the Log window. Output from SAS programs is viewed in the Output window and navigate and managed in the Results window.

Platforms on which the Software Can Be Run.

Version 7 of the SAS System is available as production on the following platforms:

- Microsoft Windows:
  Windows 95 (Build 950 or greater)
  Windows 98 (Build 1998)
  Windows NT 4.0 (Build 1381: Service Pack 3),
  Windows NT 5.0 (in an experimental mode only)
- IBM OS/2® Warp 3.0, Warp 4.0
- IBM AIX® 4.2, 4.3
- HP HP-UX 10.20, 11.0
- Sun Solaris 2.6
- Digital UNIX 4.0d
- OpenVMS Alpha 7.1
- OpenVMS VAX 6.2
- IBM OS/390® V1R1, V1R2, V1R3, V2R4
- IBM MVS 4.2
- IBM CMS 10

Availability, Pricing and Terms.

SAS Software is licensed on an annual basis. Please contact the SAS Institute directly for more information.

Contact Information

SAS Institute Inc.
SAS Campus Drive
Cary, NC 27513-2414
USA
Telephone: (919) 677-8000
Fax: (919) 677-4444
SAS Home Page: http://www.sas.com/
Statistics and Operations Research: http://www.sas.com/rnd/app/

Additional Information.

Recent papers and documentation on the survey selection and analysis procedures are available from SAS Institute's Statistics and Operations Research website at http://www.sas.com/rnd/app/da/new/dasurvey.html; see links at bottom of that page for papers and documentation.

Summary of survey software: Stata

Vendor

Stata Corporation

Types of designs that can be accommodated.

- stratified designs;
- cluster sampling;
- variance estimation for multistage sample data can be carried out through the customary between-PSU-squared-differences calculation;
- finite-population corrections can be calculated for simple random sampling without replacement of sampling units within strata.

Types of estimands and statistical analyses that can be accommodated.

- Estimation of means, totals, ratios, and proportions.
- Linear regression, logistic regression, and probit; also, tobit, interval, censored, instrumental variables, multinomial logit, ordered logit and probit, and Poisson. Point estimates, associated standard errors, confidence intervals, and design effects for the full population or subpopulations are displayed. Auxiliary commands will display all this information for linear combinations (e.g., differences) of estimators, and conduct hypothesis tests.
- Contingency tables with Rao-Scott corrections of chi-squared tests; new survey-corrected regression commands including tobit, interval, censored, instrumental variables, multinomial logit, ordered logit and probit, and Poisson.

Restrictions on number of variables or observations.

Maximum number of observations limited only by computer RAM (virtual memory can be used, but commands run slower). Maximum number of variables is 2,047 (more or less, depending on version).

Primary methods used for variance estimation.

Taylor-series linearization is used in the survey analysis commands. There are also commands for jackknife and bootstrap variance estimation, although these are not specifically oriented to survey data.

General description of the "feel" of the software.

Stata is a complete statistical software package with full statistical, data management, and graphical capabilities. It can be run interactively or in batch mode, and is fully programmable. The survey commands are part of the standard software package. Initially, data can be read in from ASCII files and a Stata-format data file created; or data in other file formats can be translated to Stata format using a stand-alone software package (Stat/Transfer or DBMS/Copy).

Platforms on which the software can be run.

- Windows (all current versions);
- Power Macintosh (OS 8.6, 9.X, or OS X);
- Alpha AXP running Digital Unix;
- HP-9000 with HP-UX;
- Intel Pentium with Linux;
- RS/6000 running AIX;

- SGI running Irix 6.5;
- SPARC running Solaris.

Software distributed as precompiled object program.

Availability, pricing and terms.

One-time purchase. Upgrade purchases are optional. Generous academic discount. Volume discounts and student discounts.

Example: academic price of one, single-user copy ranges from $369 to $625 depending on version, and includes documentation.

Contact information.

Stata Corporation
4905 Lakeway Drive
College Station, TX 77845
800-782-8272 (U.S.)
800-248-8272 (Canada)
409-696-4600 (Worldwide)
409-696-4601 (Fax)
E-mail: Stata@Stata.com
Web site: http://www.Stata.com

Additional information

This software is discussed in the review article from The Survey Statistician.

Summary of survey software: SUDAAN

Vendor

Research Triangle Institute

Types of designs that can be accommodated.

Multiple design options allow users to analyze data from stratified, cluster sample, or multistage sample designs. Sample members may have been selected with unequal probabilities, and either with or without replacement. Any number of strata and stages can be specified. In addition, different design options may be combined in one study if different sampling methods were used for parts of the population.

Types of estimands and statistical analyses that can be accommodated.

SUDAAN includes the following statistical procedures:

- MULTILOG: Fits multinomial logistic regression models to ordinal and nominal categorical data and computes hypothesis tests for model parameters. Estimates odds ratios and their 95% confidence intervals for each model parameter. Has GEE (Generalized Estimating Equation) modeling capabilities for efficient parameter estimation.
- REGRESS: Fits linear regression models to continuous outcomes and performs hypothesis tests concerning the model parameters.
- LOGISTIC: Fits logistic regression models to binary data and computes hypothesis tests for model parameters. Estimates odds ratios and their 95% confidence intervals for each model parameter.
- SURVIVAL: Fits proportional hazards (Cox regression) models to failure time data. Estimates hazard ratios and their 95% confidence intervals for each model parameter.
- CROSSTAB: Computes frequencies, percentage distributions, odds ratios, relative risks, and their standard errors (or confidence intervals) for user-specified cross-tabulations, as well as chi-square tests of independence and the Cochran-Mantel-Haenszel chi-square test for stratified two-way tables.
- DESCRIPT: Computes estimates of means, totals, proportions, percentages, geometric means, quantiles, and their standard errors. Also computes standardized estimates and tests of single degree-of-freedom contrasts among levels of a categorical variable.
- RATIO: Computes estimates and standard errors of generalized ratios of the form (Summation y) / (Summation x), where x and y are observed variables. Also computes standardized estimates and tests single-degree-of-freedom contrasts among levels of a categorical variable.
- The EFFECT statement allows users to specify contrasts of regression coefficients and hypothesis tests using simple effect names.

Restrictions on number of variables or observations.

None

Primary methods used for variance estimation.

The Taylor series linearization method (GEE for regression models) is used combined with variance estimation formulas specific to the sample design. The user does not need to develop special replicate weights since the sample design can be specified directly to the program.

Jackknife and Balanced Repeated Replication (BRR) variance estimation is also supported.

General description of the "feel" of the software.

SUDAAN uses a SAS-like language. There are two versions of Sudaan with different data interfaces:

- "SAS-callable" Sudaan: SUDAAN is called directly as a SAS procedure.
- "Standalone Sudaan": Independent program that reads external file formats, including SAS files or SPSS files.

In either case, the same programming language is used.

Platforms on which the software can be run.

- PCs under Windows 95 or later versions. This is now the primary platform for Sudaan.
- Sun SPARC computers under Solaris 2.6 and up.
- SUDAAN is distributed as a precompiled program.

Availability, pricing and terms.

Release 8.0 of Sudaan is the current version (see list of enhanced features).

SUDAAN is available under annual site licenses. Annual academic site license renewal prices range from $60 to $385 per user (depending on volume), with new licenses costing about twice as much. Government and commercial annual site license prices range from $65 to $735 per user depending on product and volume See pricing details at Sudaan Web site.

Contact information.

SUDAAN Product Coordinator
Research Triangle Institute
3040 Cornwallis Road
Research Triangle Park NC 27709-2194
Telephone: 919-541-6602
FAX: 919-541-7431
Email: SUDAAN@rti.org
URL: http://www.rti.org/sudaan/

Additional information.

SUDAAN offers public 2-day or 3-day training classes several times each year. Classes can also be taught at user sites.

The following papers about Sudaan are available on-line:
Full help manual may be viewed on-line.
Bieler and Williams (1996), "Application of the SUDAAN Software Package to Clustered Data Problems: Pharmaceutical Research."
"Analyzing Repeated Measures and Cluster-Correlated Data Using SUDAAN Release 7.5" (1997).
"Analyzing Survey Data Using SUDAAN Release 7.5" (1997, compares Taylor series, jackknife and BRR variance estimates)

An extensive on-line help library is included for interactive use.

This software is discussed in the review article from *The Survey Statistician*.
See also Shah and Barnwell (1993), "Recent developments and future plans for SUDAAN" in *Proceedings of the Survey Research Methods Section, ASA,* 657-661.